POPULAR SCIENCE

Want More?

We're told to fear robots. But why do we think they'll turn on us?

The robot uprising is a myth.

By STEVEN PINKER FEBRUARY 13, 2018

Despite the gory headlines, objective data show that people all over the world are, on average, living longer, contracting fewer diseases, eating more food, spending more time in school, getting access to more culture, and becoming less likely to be killed in a war, murder, or an accident. Yet despair springs eternal. When pessimists are forced to concede that life has been getting better and better for more and more people, they have a retort at the ready. We are cheerfully hurtling toward a catastrophe, they say, like the man who fell off the roof and said, “So far so good” as he passed each floor. Or we are playing Russian roulette, and the deadly odds are bound to catch up to us. Or we will be blindsided by a black swan, a four-sigma event far along the tail of the statistical distribution of hazards, with low odds but calamitous harm.

For half a century, the four horsemen of the modern apocalypse have been overpopulation, resource shortages, pollution, and nuclear war. They have recently been joined by a cavalry of more-exotic knights: nanobots that will engulf us, robots that will enslave us, artificial intelligence that will turn us into raw materials, and Bulgarian teenagers who will brew a genocidal virus or take down the internet from their bedrooms.

The sentinels for the familiar horsemen tended to be romantics and Luddites. But those who warn of the higher-tech dangers are often scientists and technologists who have deployed their ingenuity to identify ever more ways in which the world will soon end. In 2003, astrophysicist Martin Rees published a book entitled Our Final Hour, in which he warned that “humankind is potentially the maker of its own demise,” and laid out some dozen ways in which we have “endangered the future of the entire universe.” For example, experiments in particle colliders could create a black hole that would annihilate Earth, or a “strangelet” of compressed quarks that would cause all matter in the cosmos to bind to it and disappear. Rees tapped a rich vein of catastrophism. The book’s Amazon page notes, “Customers who viewed this item also viewed Global Catastrophic Risks; Our Final Invention: Artificial Intelligence and the End of the Human Era; The End: What Science and Religion Tell Us About the Apocalypse; and World War Z: An Oral History of the Zombie War.” Techno-philanthropists have bankrolled research institutes dedicated to discovering new existential threats and figuring out how to save the world from them, including the Future of Humanity Institute, the Future of Life Institute, the Center for the Study of Existential Risk, and the Global Catastrophic Risk Institute.

How should we think about the existential threats that lurk behind our incremental progress? No one can prophesy that a cataclysm will never happen, and this writing contains no such assurance. Climate change and nuclear war in particular are serious global challenges. Though they are unsolved, they are solvable, and road maps have been laid out for long-term decarbonization and denuclearization. These processes are well underway. The world has been emitting less carbon dioxide per dollar of gross domestic product, and the world’s nuclear arsenal has been reduced by 85 percent. Of course, though to avert possible catastrophes, they must be pushed all the way to zero.

ON TOP OF THESE REAL CHALLENGES, though, are scenarios that are more dubious. Several technology commentators have speculated about a danger that we will be subjugated, intentionally or accidentally, by artificial intelligence (AI), a disaster sometimes called the Robopocalypse and commonly illustrated with stills from the Terminator movies. Several smart people take it seriously (if a bit hypocritically). Elon Musk, whose company makes artificially intelligent self-driving cars, called the technology “more dangerous than nukes.” Stephen Hawking, speaking through his artificially intelligent synthesizer, warned that it could “spell the end of the human race.” But among the smart people who aren’t losing sleep are most experts in artificial intelligence and most experts in human intelligence.

The Robopocalypse is based on a muzzy conception of intelligence that owes more to the Great Chain of Being and a Nietzschean will to power than to a modern scientific understanding. In this conception, intelligence is an all-powerful, wish-granting potion that agents possess in different amounts.

Humans have more of it than animals, and an artificially intelligent computer or robot of the future (“an AI,” in the new count-noun usage) will have more of it than humans. Since we humans have used our moderate endowment to domesticate or exterminate less well-endowed animals (and since technologically advanced societies have enslaved or annihilated technologically primitive ones), it follows that a super-smart AI would do the same to us. Since an AI will think millions of times faster than we do, and use its super-intelligence to recursively improve its superintelligence (a scenario sometimes called “foom,” after the comic-book sound effect), from the instant it is turned on, we will be powerless to stop it.

But the scenario makes about as much sense as the worry that since jet planes have surpassed the flying ability of eagles, someday they will swoop out of the sky and seize our cattle. The first fallacy is a confusion of intelligence with motivation—of beliefs with desires, inferences with goals, thinking with wanting. Even if we did invent superhumanly intelligent robots, why would they want to enslave their masters or take over the world? Intelligence is the ability to deploy novel means to attain a goal. But the goals are extraneous to the intelligence: Being smart is not the same as wanting something. It just so happens that the intelligence in one system, Homo sapiens, is a product of Darwinian natural selection, an inherently competitive process. In the brains of that species, reasoning comes bundled (to varying degrees in different specimens) with goals such as dominating rivals and amassing resources. But it’s a mistake to confuse a circuit in the limbic brain of a certain species of primate with the very nature of intelligence. An artificially intelligent system that was designed rather than evolved could just as easily think like shmoos, the blobby altruists in Al Capp’s comic strip Li’l Abner, who deploy their considerable ingenuity to barbecue themselves for the benefit of human eaters. There is no law of complex systems that says intelligent agents must turn into ruthless conquistadors.

The second fallacy is to think of intelligence as a boundless continuum of potency, a miraculous elixir with the power to solve any problem, attain any goal. The fallacy leads to nonsensical questions like when an AI will “exceed human-level intelligence,” and to the image of an ultimate “Artificial General Intelligence” (AGI) with God-like omniscience and omnipotence. Intelligence is a contraption of gadgets: software modules that acquire, or are programmed with, knowledge of how to pursue various goals in various domains. People are equipped to find food, win friends and influence people, charm prospective mates, bring up children, move around in the world, and pursue other human obsessions and pastimes. Computers may be programmed to take on some of these problems (like recognizing faces), not to bother with others (like charming mates), and to take on still other problems that humans can’t solve (like simulating the climate or sorting millions of accounting records).

Each system is an idiot savant, with little ability to leap to problems it was not set up to solve."

The problems are different, and the kinds of knowledge needed to solve them are different. Unlike Laplace’s demon, the mythical being that knows the location and momentum of every particle in the universe and feeds them into equations for physical laws to calculate the state of everything at any time in the future, a real-life knower has to acquire information about the messy world of objects and people by engaging with it one domain

at a time. Understanding does not obey Moore's Law: Knowledge is acquired by formulating explanations and testing them against reality, not by running an algorithm faster and faster. Devouring the information on the internet will not confer omniscience either: Big data is still finite data, and the universe of knowledge is infinite.

For these reasons, many AI researchers are annoyed by the latest round of hype (the perennial bane of AI), which has misled observers into thinking that Artificial General Intelligence is just around the corner. As far as I know, there are no projects to build an AGI, not just because it would be commercially dubious, but also because the concept is barely coherent. The 2010s have, to be sure, brought us systems that can drive cars, caption photographs, recognize speech, and beat humans at Jeopardy!, Go, and Atari computer games.

But the advances have not come from a better understanding of the workings of intelligence but from the brute-force power of faster chips and bigger data, which allow the programs to be trained on millions of examples and generalize to similar new ones. Each system is an idiot savant, with little ability to leap to problems it was not set up to solve, and a brittle mastery of those it was. A photo-captioning program labels an impending plane crash "An airplane is parked on the tarmac"; a game-playing program is flummoxed by the slightest change in the scoring rules. Though the programs will surely get better, there are no signs of foom. Nor have any of these programs made a move toward taking over the lab or enslaving their programmers.

Even if an AGI tried to exercise a will to power, without the cooperation of humans, it would remain an impotent brain in a vat. The computer scientist Ramez Naam deflates the bubbles surrounding foom, a technological singularity, and exponential self-improvement:

Imagine you are a super-intelligent AI running on some sort of microprocessor (or perhaps, millions of such microprocessors). In an instant, you come up with a design for an even faster, more powerful microprocessor you can run on. Now…drat! You have to actually manufacture those microprocessors. And those [fabrication plants] take tremendous energy, they take the input of materials imported from all around the world, they take highly controlled internal environments that require airlocks, filters, and all sorts of specialized equipment to maintain, and so on. All of this takes time and energy to acquire, transport, integrate, build housing for, build power plants for, test, and manufacture. The real world has gotten in the way of your upward spiral of self-transcendence.

The real world gets in the way of many digital apocalypses. When HAL gets uppity, Dave disables it with a screwdriver, leaving it pathetically singing "A Bicycle Built for Two" to itself. Of course, one can always imagine a Doomsday Computer that is malevolent, universally empowered, always on, and tamper-proof. The way to deal with this threat is straightforward: Don't build one.

As the prospect of evil robots started to seem too kitschy to take seriously, a new digital apocalypse was spotted by the existential guardians. This storyline is based not on Frankenstein or the Golem but on the Genie granting us three wishes, the third of which is needed to undo the first two, and on King Midas ruing his ability to turn everything he touches into gold, including his food and his family. The danger, sometimes called the Value Alignment Problem, is that we might give an AI a goal, and then helplessly stand by as it relentlessly and literal-mindedly implemented its interpretation of that goal, the rest of our interests be damned. If we gave an AI the goal of maintaining the water level behind a dam, it might flood a town, not caring about the people who drowned. If we gave it the goal of making paper clips, it might turn all the matter in the reachable universe into paper clips, including our possessions and bodies. If we asked it to maximize human happiness, it might implant us all with intravenous dopamine drips, or rewire our brains so we were happiest sitting in jars, or, if it had been trained on the concept of happiness with pictures of smiling faces, tile the galaxy with trillions of nanoscopic pictures of smiley-faces.

I am not making these up. These are the scenarios that supposedly illustrate the existential threat to the human species of advanced artificial intelligence. They are, fortunately, self-refuting. They depend on the premises that 1) humans are so gifted that they can design an omniscient and omnipotent AI, yet so moronic that they would give it control of the universe without testing how it works; and 2) the AI would be so brilliant that it could figure out how to transmute elements and rewire brains, yet so imbecilic that it would wreak havoc based on elementary blunders of misunderstanding. The ability to choose an action that best satisfies conflicting goals is not an add-on to intelligence that engineers might slap themselves in the forehead for forgetting to install; it is

intelligence. So is the ability to interpret the intentions of a language user in context. Only on a television comedy like Get Smart does a robot respond to “Grab the waiter” by hefting the maitre d’ over his head, or “Kill the light” by pulling out a pistol and shooting it.

More technology stories:

When we put aside fantasies like foom, digital megalomania, instant omniscience, and perfect control of every molecule in the universe, artificial intelligence is like any other technology. It is developed incrementally, designed to satisfy multiple conditions, tested before it is implemented, and constantly tweaked for efficacy and safety. As AI expert Stuart Russell puts it: “No one in civil engineering talks about ‘building bridges that don’t fall down.’ They just call it ‘building bridges.’” Likewise, he notes, AI that is beneficial rather than dangerous is simply AI.

Artificial intelligence, to be sure, poses the more mundane challenge of what to do about the people whose jobs are eliminated by automation. But the jobs won’t be eliminated that quickly. The observation of a 1965 report from NASA still holds: “Man is the lowest-cost, 150-pound, nonlinear, all-purpose computer system that can be mass-produced by unskilled labor.” Driving a car is an easier engineering problem than unloading a dishwasher, running an errand, or changing a diaper, and at the time of this writing, we’re still not ready to loose self-driving cars on city streets. Until the day battalions of robots are inoculating children and building schools in the developing world, or for that matter, building infrastructure and caring for the aged in ours, there will be plenty of work to be done. The same kind of ingenuity that has been applied to the design of software and robots could be applied to the design of government and private-sector policies that match idle hands with undone work.

Adapted from ENLIGHTENMENT NOW: The Case for Reason, Science, Humanism, and Progress by Steven Pinker, published by Viking, an imprint of Penguin Publishing Group, a division of Penguin Random House LLC. Copyright © 2018 by Steven Pinker.

This article was originally published in the Spring 2018 Intelligence issue of Popular Science.

Latest News

Want more news like this?

Sign up to receive our email newsletter and never miss an update!

Enter email address
Sign Up

By submitting above, you agree to our privacy policy.